DOCUMENT RESUME

ED 255 543 TM 850 122

AUTHOR	Carlson, Sybil B.; Camp, Roberta
TITLE	Relationships between Direct and Indirect Measures of Writing Ability.
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	1 Apr 85
NOTE	23p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 31-April 4, 1985).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	College Entrance Examinations; Computer Assisted Testing; *English (Second Language); *Essay Tests; Experimenter Characteristics; Higher Education; Holistic Evaluation; Interrater Reliability; Language Proficiency; Scoring; Test Construction; *Test Reliability; *Test Validity; *Writing Evaluation; Writing Skills
IDENTIFIERS	Graduate Record Examinations; Test of English as a Foreign Language; Writers Workbench (Computer Software)

ABSTRACT

This paper reports on Educational Testing Service research studies investigating the parameters critical to reliability and validity in both the direct and indirect writing ability assessment of higher education applicants. The studies involved: (1) formulating an operational definition of writing competence; (2) designing and pretesting writing assessment instruments; (3) identifying and controlling parameters influencing a writing assessment program; and (4) the scoring and interpretation of measurement outcomes. Data from 638 college applicants representing three foreign language groups plus a small sample of native English speakers were collected from four writing samples scored by several methods, Test of English as a Foreign Language (TOEFL) scores, and Graduate Record Examination (GRE) General test scores. Correlational and factor analyses revealed these key findings: (1) holistic scores, discourse-level scores, and sentence-level scores were so closely related that holistic scores alone should be sufficient; (2) correlations among topics were as high across topic types as within topic types; (3) similar scoring by different types of raters suggest substantial agreement in the standards used; (4) writing sample and TOEFL scores were highly related but each also measured other aspects of English proficiency; and (5) relationship patterns between holistic writing sample scores and item type scores within GRE General Test sections were consistent with other GRE studies. (Author/BS)

Presentation--Session B2
NCME Annual Meeting
April 1, 1985

Relationships Between Direct and Indirect Measures of Writing Ability

Sybil B. Carlson and Roberta Camp, Educational Testing Service

A series of research studies conducted at Educational Testing Service have been investigating the parameters critical to reliability and validity in the direct and indirect assessment of the writing ability of applicants to higher education. Although the studies are based on performance of students taking the Test of English as a Foreign Language (TOEFL) and the Graduate Record Examinations (GRE) General test, they have yielded data that have general implications for the design of writing assessment programs at the postsecondary level.

The results of these investigations have direct implications for the assessment of writing performance at the postsecondary level. In that evaluation serves as a tool to facilitate various forms of educational decision making, the valid, objective measurement of writing performance must be perceived as one of the fundamental keys to attaining educational objectives. As we develop and evaluate measures of writing competence, several measurement issues require careful attention. This study focuses on, and has collected additional information regarding the following issues: the formulation of an operational definition of writing competence within functional contexts, the design and pretesting of appropriate instruments to reflect the intended purposes of evaluation, the identification and control of parameters that influence a writing assessment program, and the scoring and interpretation of measurement outcomes from the perspective of various forms of inference regarding validity and reliability.

Numerous variables influence the outcomes and interpretations of writing assessment programs. When assessment programs take into account the knowledge and experience gained in the field, they can be systematically designed to (1) build in controls over variables that can be controlled, and (2) identify variables that cannot be controlled effectively. This research involved very direct and far-reaching applications, in that the TOEFL program has been seriously considering the addition of a writing component to the TOEFL, and the GRE program has continued over the years to support efforts to assess the validity and interpretation of GRE scores from different perspectives. Thus, in addition to implementing a solid research design, the study integrated components of the real-world context in which writing skills would be viewed in relation to the two testing programs.

Funding provided by the GRE and TOEFL programs afforded the opportunity for us to develop a model for the design of a valid and reliable writing assessment program. Our paramount objective was to test the effectiveness of a model that would be useful to other writing assessment programs, both at ETS and in postsecondary institutions outside of the organization. In each stage of the investigation, we paid careful attention to detail, since any one element could critically influence the others. Specifically, the investigations have provided insights into the connections and comparisons among the following: (1) criteria for writing across academic disciplines, (2) EPL and ESL performance (including three major language groups), (3) perforrance across discourse modes, (4) iuformation obtained from different methods of scoring, and (5) results from direct and indirect measures of writing ability.

The primary purpose of this.study was to determine the relationships of TOEFL and GRE General Test scores with the kinds of writing tasks that first-year students are expected to perform. These data provide important information regarding the construct validity of the GRE and the TOEFL, information which should be useful to those who interpret the scores from these tests as well as to ETS test developers who may be considering the addition of direct measures of writing ability, in the case of the TOEFL, and of indirect and direct measures of writing ability for future GRE forms. The study involved the collection of four writing samples from native and nonnative speakers of English who were seeking admission to undergraduate and graduate levels of education in the United States and Canada. In addition, recent GRE General and TOEFL scores were obtained for the appropriate groups of candidates (e.g., candidates for admission to undergraduate programs do not take the GRE). The TOEFL scores include an indirect measure of writing skill, the Structure and Written Expression section of the TOEFL; scores on a comparable indirect measure (a section of a retired form of the LSAT) were obtained for native-speaking GRE candidates. The standardized test scores then were related to holistic and analytic scores on the writing samples.

The most significant and fundamental tasks for this research required (1) the design of writing assessment instruments and (2) the collection and scoring of writing samples with these instruments. Elaborate planning was necessary, since the validity and usefulness of the information gained by the data analyses would depend on the quality of the measurement process. In order to achieve the best and most appropriate assessment of writing skills, the study design took into account the numerous perspectives that the "state of the art" in the evaluation of writing ability has to offer. We combined the knowledge and experience accumulated by a variety of disciplines--writing assessment and instruction, psychological measurement, linguistics, contrastive rhetoric, and instruction in English as a second language (ESL). Each of these fields offer insights garnered from theory, research, and practice. Our first planning objective focused on the definition of competence in writing, a definition that emphasizes the situational context of writing assessment appropriate to the objectives of the TOEFL and the GRE General Test as indicators of a student's ability to write English. This definition was formulated on the basis of information drawn from the areas of writing assessment, communicative competency, and contrastive rhetoric. Our second planning objective required the design of a validation study that depended on the development of effective instruments to evaluate written competence and on rigorously implemented data collection and scoring procedures. The following section briefly summarizes the framework for formulating a functional definition of writing ability, including our survey of academic writing skills. The subsequent sections describe the design and implementation of the validation study.

A Definition of Writing Competence

The first investigation was designed to identify and describe operationally the expectations of writing competence required of native and nonnative speakers of English at the beginning of their postsecondary education. The information gathered took into account the various factors that should be considered in defining communicative competence in writing--the functional task demands for which students are expected to be prepared, as well as the perceptions, sometimes culturally influenced, of those who evaluate them. Since the research was conducted for testing programs that provide indicators of readiness to participate in an English-based curriculum at the undergraduate and graduate levels, the objectives of the writing assessment study reflect the intended measurement purposes of the tests. These specific operational definitions subsequently would be reflected in the development of instruments to assess functional communicative writing competence in the context for which the instruments were intended.

The TOEFL Survey of Academic Writing Tasks

The literature on functional communicative competency served as the basis for the design of a research project that would provide a definition of writing task demands in postsecondary academic settings. The primary objective of this project (Bridgeman and Carlson, 1983) was to identify and describe operationally the expectations of writing competence required of nonnative speakers of English at the beginning of their educational experiences in institutions of higher education in the United States and Canada.

The survey questionnaire was completed by faculty in 190 academic departments at 34 universities in the United States and Canada with high foreign student enrollments. At the graduate level, six academic disciplines with relatively high numbers of nonnative students were surveyed: business management (MBA), civil engineering, electrical engineering, psychology, chemistry, and computer science. Undergraduate English departments were chosen to document the skills needed by undergraduate students.

The detailed findings are described in another publication (Bridgeman & Carlson, 1984). Briefly summmarized, the faculty members surveyed appear to view student writing skills from the standpoint of functional communicative competencies. For example, the written products prepared by students in different disciplines may be considered competent to the extent that they meet the task demands--particularly kinds of writing assignments and certain skills--that are specific to a discipline. In addition, faculty members reported that written assignments are evaluated on the basis of discourse-level characteristics, rather than word- or sentence-level characteristics, and that they perceived the discourse-level writing skills of natives and nonnatives to be fairly similar. Grammatical competency, however, tends to influence evaluations of student writing to some extent, in that respondents reported that nonnatives are more deficient in word- and sentence-level skills than are natives.

The survey indicated that no single essay topic type was universally accepted by all of the academic disciplines surveyed. In the multidimensional scaling, Types H (Describe and Interpret a Chart or Graph) and E (Compare and Contrast) were further apart in the space than any other pair of types, suggesting that they were perceived as distinctly different tasks. Thus the Type E topic type was selected to serve as an effective contrast to the Type H topic type; since departments perceived these two types as distinctly different, it seemed likely that writing samples elicited by Types H and E elicited different writing skills, as well.

The Validation Study

The purpose of the subsequent study, designed on the basis of the survey results, investigated whether scores on the TOEFL and the GRE General test, as well as scores on indirect measures of writing ability, are predictive of writing skills on the kinds of tasks faced by first-year undergraduate and graduate students. This investigation carefully controlled four primary factors critical to the various forms of validity and reliability in the direct assessment of writing ability--task, administration and data collection, scoring method, and psychometric and interpretation factors.

Instrument Development

In preparation for making comparisons of direct measures of writing ability with indirect measures and with TOEFL and GRE scores, writing tasks were carefully designed for the assessment of performance on writing samples. The task factors, which are presented by the writing stimulus or prompt, consist of the following: determination of content, development of the stimulus material, and pretesting of writing stimuli to determine whether the expected performance is elicited and can be evaluated.[1] For the readers of the student papers, who represented two different disciplines--ESL and English composition--a questionnaire was developed to survey the readers' general perspectives on the evaluation of writing and their reactions to the scoring of the writing samples in this study.

[1] A full discussion of these task factors will appear in the chapter, "Testing ESL Writers" (Carlson and Bridgeman, in press).

The process of instrument development, a critical element of the study, demanded attention to, and, to the extent possible, control of the numerous factors that influence a direct assessment of writing ability. Besides heeding the many considerations that normally influence the design of the writing task, we needed, through pretesting and pilot testing of topics, to test our assumptions concerning the writing performance that would actually be elicited by our particular topics and the tasks they presented. The topics then were pretested, resulting in the selection of a reduced number of topics with the potential to tap writing performance effectively. Furthermore, these topics were pilot tested, and the resulting writing samples were scored in an essay reading that focused on the writing performance elicited by the topics. Eventually, the final topics that were selected for administration to the large sample of international and U.S. students were refined and formatted in carefully designed test booklets.

Our survey of academic writing tasks provided the basis for the development of writing assessment instruments. The survey enabled us to define writing competence functionally in terms of the writing tasks that beginning postsecondary students would be expected to perform, and the measurement objectives of the TOEFL and GRE General tests. In addition, the survey guided us in the selection and implementation of the parameters influencing the measurement of writing skills, such as specific approaches to scoring, that were critical to this writing assessment data cc lection.

For this project, we therefore developed two topics of Type H (Chart/Graph) and two topics of Type E (Compare/Contrast) to which each student would respond. In order to administer topics that would most effectively meet the measurement objectives of the study, several topics of each type were developed and pretested.

The pretesting and pilot testing readings of the papers enabled us to determine how well the specific topics actually elicited the kinds of academic writing skills we intended to evaluate. We concluded that the writing tasks were successful in eliciting the targeted performances which corresponded to our definition of academic writing competence with respect to the four topics. The readings also afforded the opportunity to identify and describe writing behaviors and the criteria for their evaluation that would be used in assisting readers when they established criteria for scoring papers during the formal reading sessions.

Test Administration Factors

Major factors in test administration that contribute to the outcomes of writing assessment were taken into consideration:

- The physical layout of the writing stimulus was designed to give writers the opportunity for prewriting tasks of planning and organization and to suggest the expected length of the writing sample.

- Directions to administrators were designed to minimize such adverse conditions in the testing room as uncomfortable temperature, poor lighting, noise, and poor writing surface.

These factors are critical to any testing situation but assume greater importance when students are asked to generate and produce written responses.

Data Collection Factors

Each native and nonnative English-speaking candidate sample produced four writing samples, two samples per topic type. We collected this number of samples in order to elicit a reasonable representation of writing skills, as well as an indication of the degree of consistency in the performance of individuals across similar and different tasks.

Sample. We obtained a total sample of 660 candidates for undergraduate and graduate study representing three language groups (Spanish, Arabic, and Chinese[2]) plus a group of native-English-speaking grpduate students from the United States. The group of students applying for admission at the graduate level was further subdivided into two major field categories: "hard" science, and social science/humanities (including business).

Testing procedures. Because languag, skills can change dramatically in a relatively short period of time, testing students in the United States some months after they took the TOEFL in their native countries might lead to inexplicable confounding and uninterpretable results. Instead, we tested students at foreign centers as close in time as possible to when they took the TOEFL. GRE scores should be less subject to short-term fluctuations, and any student who had taken the GRE up to six months before the TOEFL or who was scheduled to take the GRE up to six months after the TOEFL was eligible for inclusion in the sample.

The international centers were selected, with the assistance of TOEFL program staff, based on the following criteria: having candidates from the desired language groups, having candidates representing diverse ability levels, having a reasonable balance of undergraduate and graduate candidates, and having substantial numbers of GRE (recent past or potential) candidates.

[2] Most Chinese TOEFL candidates are from Taiwan, but few undergraduates are tested in Taiwan. Large numbers of Chinese candidates for admission as undergraduates come from Hong Kong. Thus, we anticipated that Chinese graduate candidates would be drawn from test sites in Taiwan, and undergraduates from Hong Kong. This would provide for the greatest generalizability of the results to the actual TOEFL population. However, given the known differences between education in Taiwan and Hong Kong, the confounding of location with undergraduate status must be considered when the results are interpreted.

Writing samples from GRE candidates in the domestic sample were collected during special testing sessions at five major university testing centers after we had identified and selected recent GRE General Test takers. Since the GRE General Test does not contain an indirect measure of writing skills, the GRE candidates at domestic sites also took a brief objective test of writing skills, a retired form of a test of writing skills formerly used by the Law School Admissions Testing program. Thus we were able to compare indirect measures with direct measures of writing for the native GRE cand aates, as well as for the nonnative TOEFL and GRE candidates.

Development of Scoring Methods for the Direct Assessment of Writing

The data collection resulted in a total of 660 booklets, each containing four complete writing samples, a total of 2,640 papers. The four topics had been assembled in the booklets in eight different sequences, in order to control for order effects.

In order to compare the results of scoring the writing samples by using different scoring methods, the papers were scored in several ways, as follows:

- The holistic scoring of all booklets, primarily on the first day of the essay reading weekend
- The two-score scoring of all booklets, for Discourse/Sentence characteristics, primarily on the second day of the essay reading weekend
- The holistic scoring of a representative subsample of the papers by subject matter experts in two major fields of graduate education
- The "analytic" scoring of the features of a representative subsample of the papers using the Writer's Workbench software

Scoring methods. Selection of an appropriate scoring method for a writing sample depends on the purposes of the assessment. A holistic evaluation (i.e., a single score representing the overall impression created by the sample) may be more efficient for making selection or placement decisions, whereas a more analytic framework (i.e., separate scores for a number of different organizational and grammatical features of the sample) may be more useful for providing diagnostic information to teachers. Although other methods (e.g., error counts) may yield more objective scores as a rough index of second language proficiency, they may be poor indicators of functional communicative competence.

Holistic scoring is impressionistic, but it is not haphazard. Considerable care must go into selecting sample essays (range finders) that represent each point on the score scale, and thorough training of the readers is necessary. Such training involves discussion among the readers to reach consensus on the criteria. During a reading session, continual checks must be made to ensure that no reader is straying from the standards originally set. Since the scorer judgments are subjective, each essay should receive at least two independent readings. The scores from the two readers are typically added together to form the single holistic score.

Holistic evaluations may be influenced by a number of features of an essay, including content, organization, sentence structure, and mechanics (Freedman, 1979; Breland & Jones, 1982). If a single holistic score is to be used, the raters must agree on how to score essays that present a large discrepancy between organizational and mechanical skill. They must also agree on which mechanical errors are most serious. This judgment of error gravity may stem from a strictly functional communication point of view (Does this error interfere with what the author is trying to say?), or it also may penalize errors that are stylistically undesirable (e.g., redundancy, run-on-sentences). In addition, raters must agree on how to evaluate essays that contain complex sentence structures, and in which the writers make errors in trying to write complex sentences, versus essays that use only simple sentences, but contain few errors. In her research, Greenberg (1983) noted that ability to avoid errors predicted teachers' quality ratings better than the writer's ability to handle complex syntactic structures. She found that one major problem consisted of word form errors. Shaughnessy (1977), in fact, recognized that word form errors exemplify "advanced errors." Such errors indicate attempts to acquire formal academic vocabulary, in spite of the risk of making errors. Thus more competent writers may commit more errors, yet may be penalized by raters who focus on the lack of errors as a predominant feature of good writing. During the training for holistic scoring, discussion about errors should be limited in order not to interfere with the process of reading for total impression, and to ensure that particular features of writing do not unduly influence that total impression.

Despite the most rigorous procedures in the training of scorers, holistic scoring schemes inevitably require some degree of subjective judgment, and these subjective judgments may be particularly difficult when the writer and reader (scorer) do not share a common set of cultural conventions and expectations. These conventions go far beyond mere differences in grammatical rules. The work of Robert Kaplan (1966) clearly demonstrated cultural differences in patterns of logic used to order ideas within paragraphs. Thompson-Panos and Thomas-Ruzic (1983) recently noted certain contrasting features of English and written Arabic that may contribute to perceived weaknesses in the writing of Arab ESL students. For example, paragraph development in Arabic languages consists of a series of parallel constructions connected by coordinating conjunctions, thus de-emphasizing the use of subordination that is valued in English paragraph organization.

ESL teachers who are aware of distinct cultural patterns may assign essay ratings that differ significantly from ratings of English teachers with no ESL experience. On the other hand, if the criterion for competence is success in a standard course in a United States university, the "insensitive" ratings may better predict academic performance than the culturally sensitive ratings. In this study, we compared ratings by ESL readers with ratings by readers whose predominant experience is with native speakers of English. In addition, these ratings were compared to ratings given by faculty members in engineering and in the social sciences. The classic research of Diederich, French, and Carlton (1961) suggests that even among native speakers different "schools of thought" exist among readers, and that certain professions are more likely to emphasize a particular characteristic. For example, lawyers appear to focus more on organization, whereas editors tend to focus on style and wording. In our research, the essay readers completed a questionnaire intended to identify the features they attend to when evaluating a composition.

Because analytic scoring yields more scores than holistic scoring, it is potentially more valuable for prescribing educational interventions for individual students. One scoring scheme that has been used extensively with ESL students provides separate scores for content, organization, vocabulary, language usage, and mechanics (Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey, 1981). Other analytic scoring schemes provide for even finer-grained analysis. However, the apparent advantage of several separate scores is frequently an illusion; the reader's general impression is likely to influence ratings on each of the "separate" aspects being evaluated. In addition, analytic ratings are very time consuming. Wiseman (1969) found that four general impression markings were equivalent in time and effort to one analytic marking. As noted previously, despite considerations of efficiency, a single holistic score may not adequately describe an ESL student with discrepant organizational and mechanical skills. Further research is needed to determine the best compromise between a single score and a complex analytic scoring scheme, as well as which kinds of scores are more appropriate to specific situational contexts.

The most promising means for the objective scoring of essays may be by computer software such as Bell Labs' "Writers Workbench" (Cherry, 1983; Kiefer, et al., 1983). This sophisticated word processing tool can identify such features as spelling errors, overuse of a particular word, and sentences that are consistently too long or too short. Analysis of these structural features might help some writers to improve their writing. However, this kind of computer program cannot judge how well a piece of writing accomplishes its main purpose of communicating with its intended audience, nor can it evaluate features such as development and organization. The subjective impression of coherence that a reader "receives" from the written communication cannot be duplicated by a mechanical count of cohesive elements (Carrell, 1982).

Scorers. The scorers included individuals experienced with assessment in ESL and English composition, including a core of scorers experienced in holistic scoring.

In order to obtain additional and independent scores for the writing samples, we also obtained ratings for a subsample of papers from faculty members from the two academic disciplines with the largest foreign student enrollments. They were asked to evaluate the papers from the perspective of writing skills exhibited at the time of admission to their program, rather than from the perspective of writing skills expected to be developed as students develop discipline-specific writing skills. These ratings by faculty members made it possible to compare scores assigned by subject matter experts with scores assigned by writing experts, providing some indication of the extent to which points of view regarding writing competence reflect different perspectives within these disciplines. Four faculty members in each of two disciplines, the social sciences and hard sciences, assigned ratings to representative samples of papers written in response to two essay topics, one of each of the two types.

The reliability and validity of methods used for scoring of writing samples is influenced strongly by the readers who apply these methods. When high interreader reliability coefficients are obtained, this may have two explanations: (1) the training sessions enabled readers, who may or may not have common views, to agree on common criteria for evaluation, and/or (2) despite the training, readers, especially those who are involved in the field of writing (English or ESL), tend to agree on criteria for evaluation. With low or moderate interreader reliability coefficients, other questions are raised with regard to the following: (1) how and if the training sessions could have been improved to obtain higher agreement; (2) what readers perceive to be their personal criteria for good writing; and (3) whether the personal criteria held by readers are significant to the evaluation of writing and should have been taken into account during the training. In order to gain information about readers' points of view, a reader questionnaire was designed. The staff decided that readers would be asked to repond to the instrument at the conclusion of each day of essay reading, rather than prior to the readings, to avoid heightening reader sensitivities with questionnaire prompts about evaluation criteria. The reader questionnaires differed slightly on the two reading days because readers would be asked to react to and compare the two different scoring methods used during each session at the conclusion of the second day of reading.

Psychometric and Interpretation Factors

Although psychometric considerations of reliability and validity are essentially the same for ESL essays as for essays written by native speakers, the unique cultural and linguistic characteristics of ESL students require special attention.

Reliability or consistency of essay scores can be assessed in a number of different ways (intrarater, interrater, across topics within genre, across genre). Intrarater reliability indicates how consistent a single rater is in scoring the same set of essays twice with a specified time interval between the first and second scoring. Interrater reliability estimates the extent to which two or more raters agree on the score that should be assigned to an essay. When essay writers and raters represent different cultural perspectives, interrater reliability is likely to be lower than when both essay writers and raters come from a homogeneous group. But even if interrater reliability is perfect, the claim cannot be made that the essay test is perfectly reliable. Other factors such as variations over time, from one topic to another, and from one sample of students to another also must be considered.

Intertopic reliability assesses the extent to which the rank ordering of student scores depends on the topic. Scores will vary from one topic to another even within the same general topic type (e.g., compare and contrast).

High reliability does not provide sufficient evidence that a test is valid. Instead, the test may be measuring consistently a variable that is not the criterion of primary interest. Thus, a 30-minute writing sample might be judged reliable, but it might not serve as a valid indicator of the student's ability to write a long paper without limitations on time and with an opportunity to make extensive initial drafts. As Cronbach (1971) has noted, it is not tests that are validated but rather interpretations of data from tests used in specific contexts. Scores from an essay test may be valid for one purpose but not another. For instance, a test that serves as a valid indicator of skill in writing a narrative essay may have little value in predicting a student's ability to meet the writing demands in a graduate engineering program. Furthermore, a test that is considered a valid predictor of success in meeting the writing demands of undergraduate study for native speakers may or may not predict with comparable validity for ESL students.

Optimally, validity should be determined by establishing that a test is measuring the same performance objective that a good external criterion also is measuring. When the parameter that condition a measure of writing skills are ta[illegible] o account, the external appearance of a writing sample topic, or it[illegible] validity, is not sufficient to ensure the validity of the performan[illegible] c is intended to be measured. An objective means for determining the validity of scores on a writing sample can be achieved by correlating these scores with scores on other measures that have been demonstrated to predict well to the same criterion. This criterion, likewise, must have evidenced validity and reliability. One frequently used criterion of academic success, such as the grade point average, may not meet consistently the constraints of validity and reliability. Instead, valid and reliable scores on an established test that has been shown to predict to the criterion (i.e., grades) may serve as a more objective indicator for validating writing sample scores. The validity of scores for writing samples that are included in standardized tests, for example, is established by demonstrating that the scores are highly correlated with scores on indirect measures of writing ability.

Ideally, however, scores on direct and indirect measures would not be perfectly correlated. Because a writing sample requires the production of a composition in contrast to the recognition of correct responses on a multiple-choice test of writing ability, we would not expect the two types of test to assess identical skills. Instead, they would be highly correlated because some of the skills they are measuring overlap and reflect a form of "general" writing ability. In addition, writing samples would be expected to contribute additional information about writing performance that is not yielded by an objective test, thus explaining an imperfect correlation.

Test validation is a process of accumulating evidence to support inferences made from test scores, reflecting the value of a test for an intended purpose; more sources of evidence are better than fewer. For this study, we intentionally planned to score the writing samples in various ways, and to relate these scores to other measures, in order to obtain as much information as possible regarding the validity of direct measures of writing in the TOEFL and GRE contexts.

Data Analyses

We performed several statistical analyses of the data, consisting of correlational and factor analyses. The data analyses were conducted to reveal the degree of relationship among several variables--GRE total and subtest scores, TOEFL total and subtest scores, scores on indirect measures of writing ability (included in the TOEFL and as a separate test for native speakers of English), and the different scores derived from direct measures of writing ability. In addition, the obtained relationships were examined with respect to the different language groups (Arabic, Chinese, Spanish, and English). The objective of the data analyses was to provide information about the content and construct validity of the GRE and TOEFL examinations; in particular, the data would suggest the extent to which writing ability contributes to GRE and TOEFL test scores. The results of the data analyses are discussed thoroughly in the full report (Carlson, Bridgeman, Camp, & Waanders, 1985) of the research, which will be available in the near future as an ETS Research Report and a GRE General Report. The overall results and conclusion are summarized briefly in the following section.

Summary of Results and Conclusions

This research generated a considerable amount of information contributing to the validity of measures of English language proficiency--writing samples, the TOEFL, and the GRE General Test. A summary of the major findings follows:

o The two scoring methods for the writing samples, holistic and discourse-/sentence-level (D/S), yielded essentially the same mean levels of performance and were highly correlated, indicating that the two-score method may not provide any significant advantage over the one-score method. Aside from the high correlations among holistic and D/S scores, we observed that (1) it was very difficult to select sample papers for scoring sessions that represented reliably different values of D and S, and (2) although readers could agree on the levels of performance for D and S, they perceived the constructs of Discourse-level and Sentence-level features to be unclear and confounded (thus challenging the validity of separating judgments on the basis of D and S).

o The means of the writing samples scores reflected level differences for the three language groups for whom English is not their primary language. For every writing sample score, the means were the lowest for the Arabic sample, in the middle for the Chinese sample, and the highest for the Spanish sample.

o The mean holistic and D/S scores obtained by the sample of United States candidates on the writing samples were considerably higher than the mean scores for the foreign group, not a surprising result since the focus of the study was on measures that assess English language proficiency.

o The reliabilities of all of the scores assigned to the writing samples were remarkably high, indicating that the consistent scoring of writing samples can be achieved (under the optimal scoring conditions described in previous chapters). The various types of evidence for reliability for the holistic scores consisted of interrater reliability, reliability across topics, and reliability within language groups. For the Discourse- and Sentence-level scores, evidence for reliability consisted of interrater reliability, reliability across score types and across topics, reliability within language groups, and reliability across ESL and English readers.

o Correlations among the topics were as high across topic type as within topic type. This result suggests that (1) the different topics did not elicit qualitatively different writing performance, and/or (2) the readers maintained a comparable scale for evaluating the writing samples, despite performance fluctuations from topic to topic.

These positive results, however, should not be interpreted as evidence that papers written in response to any topic or type of topic would yield equivalent reliability. The topics were selected on the basis of previous research indicating that specific kinds of topics would serve as more appropriate stimuli to reflect the academic writing task demands experienced by students in higher education in the United States. Carefully controlled conditions of design and pretesting, and of scoring methods that emphasized functional academic English proficiency, would need to be replicated to attain similar results.

Both this study and our previous survey of academic writing tasks have demonstrated, though, that topics designed to elicit the English skills of TOEFL candidates in different disciplines do not need to be subject-specific in order to evaluate writing performance effectively as long as they are within the context of relevant academic competencies.

o Whatever differences in the perception of good writing may exist among regular English teachers, ESL teachers, social science teachers, and engineering teachers, these differences do not interfere with the ability of these diverse groups to rank students' writing samples in the same order. When subject-matter experts in engineering and the social sciences were asked to rate representative subsamples of papers written in response to two topics, the professors' ratings were highly correlated with each other--the mean social science ratings correlated .92 with the mean engineering ratings for each of the two topics. When compared with the holistic scores assigned during the regular scoring session for the compare/contrast topic (Space), the mean social science judgment correlated .86 with the holistic scores, and the mean engineering judgment, .92. For the chart/graph topic (Farming), the correlations were .83 and .82, respectively. This outcome further supports the assumption that general agreement exists, even when not formally identified and verbalized, concerning standards for academic writing competence.

These results also can be explained by two design factors: (1) the professors were instructed to evaluate the papers from the perspective of writing competence required of students to succeed in their graduate-level departments, as opposed to writing competence in general; and (2) they were supplied with a limited number and representative sample of papers such that the task was to some extent more highly structured than the task addressed by the holistic readers.

o The reader responses to the questionnaires provided information about the points of view with regard to the evaluation of writing skills and the readers' exposure to different methods of scoring papers on the same topics. Reader ratings of the features of written assignments suggested that the readers perceived that they were attending to somewhat different characteristics of writing competence during the holistic scoring than during the D/S scoring. However, although the readers may have focused on different features, the means and standard deviations of the scores indicated that the different scoring methods did not yield different score levels. Thus the evaluations of the quality of writing competence were consistent, regardless of scoring method. These results suggest that papers that are strong on one measure (D) are strong on another (S), or that perceptions of D and S go hand in hand. This finding also supports the supposition held by readers of compositions that general agreement exists, even when not formally identified and verbalized, concerning the standards for writing competence.

Data obtained from the Writer's Workbench, as a tool for investigating the features of writing samples that may be salient to readers, suggested that further investigation may provide useful information regarding relationships among features of the papers and the scores assigned to the papers.

- In response to other questions on the questionnaire, a considerable number of readers (70 percent) felt that the scores they were asked to assign during both scoring sessions were appropriate to the particular sample of papers.

- Many readers indicated that they would be very uncomfortable with attempting to assign descriptions to score levels because individual papers at one score level can differ considerably. Most readers appeared to agree, however, that sample papers at each score level could be useful and meaningful if provided in a score manual for an operational writing sample testing program, both to other readers of writing samples and to those who would interpret writing sample scores.

- A principal axes factor analysis with varimax rotations of holistic scores and TOEFL section scores resulted in a two-factor solution. The two factors appear to be method factors, one consisting of scores on the three sections of the TOEFL, and the other, of holistic scores on papers written in response to the four topics. One interpretation of the two factors suggests that performance on measures of English language proficiency becomes more differentiated when English proficiency measures require a candidate to respond by applying different cognitive processes--recognition vs. production.

- A comparison of the relationships of writing sample and TOEFL mean scores showed that the pattern of means across the three language groups is highly consistent. This lack of interaction between type of score (writing sample or multiple-choice) and language group is consistent with the notion that both types of scores may assess, to a great extent, the same underlying language proficiency dimension. However, there is some evidence that the between-groups differences are smaller for the scores on the writing samples than for the TOEFL.

- The correlations between the holistic score total (direct evidence of a productive skill) and the TOEFL total (measures of receptive skills and indirect measures of writing) indicate that the two measures evaluate English proficiency to a considerable degree, but that the overlap between the two instruments is not perfect. The writing sample contributes additional information regarding English proficiency, in that a competently executed writing sample demonstrates the application of cognitive abilities far beyond the mastery of mechanics. The objective of the TOEFL is to provide evidence of mastery of more basic English language skills, but not of higher-order writing skills such as organization and quality of ideas.

  In addition, the relationships of the writing sample score with other sections of the TOEFL are consistent with the pattern of relationships among the TOEFL sections, such as reported in previous research (Pitcher and Ra, 1967; Pike, 1976), although the sizes of the correlations obtained in this study are somewhat lower. The earlier research results, however, cannot be compared directly with our findings because of basic design differences. In the previous studies, the composition of the TOEFL was different, since it was the five-section version used prior to 1976. In addition, the topics differed considerably--topics in Pike's study included more explicit and restrictive instructions, and elicited papers written in a narrative form. Pike also investigated three native country groups (from Chile, Peru, and Japan) whereas this research targeted a different configuration of native languages (Arabic, Chinese, Spanish). The consistent pattern of relationships obtained in the three studies, however, lend further support to the validity of the TOEFL and direct measures of writing ability.

- For the foreign sample, the correlation of scores on the writing sample with scores on the TOEFL total and with the GRE-V are nearly identical, indicating that the writing sample scores serve as an indicator of English language skills. For foreign candidates, however, the GRE-V requires a high level of English proficiency in contrast to the TOEFL.

- The correlation of writing sample scores with GRE-V scores is substantially higher in the total sample than for the international sample because the United States students scored relatively high on both measures. The correlations of scores on sections of the GRE General test with the TOEFL and writing sample scores present remarkably stable patterns of relationships.

- When the holistic writing sample scores, averaged over four topics, were related to scores on item types within the sections of the GRE General test, the observed pattern of correlations

> was consistent with the relationships reported in other GRE studies. Specifically, the analytical reasoning and logical reasoning scores were not highly correlated, and the analytical reasoning items were more highly correlated with the quantitative items than were the logical reasoning items. On the other hand, the logical reasoning items were more highly correlated with the verbal items than were the analytical reasoning items. The holistic scores were more highly correlated with the logical reasoning items than with the analytical reasoning items, further indication that the holistic scores reflect verbal ability as measured by relevant item types on the GRE General test.

The results suggest that, with careful topic selection and adequate training of raters, writing samples can provide a reliable measure of the English proficiency of nonnative speakers, as well as native speakers of English, and that direct measures of writing performance, although substantially corre'ated with multiple-choice measures such as the TOEFL and GRE General test, contribute additional information regarding the English proficiency of candidates.

There was no indication of any important differences between the two topic types (chart/graph interpretation and compare/contrast) used in this study. However, it is important to remember that both topic types represent structured, academically oriented writing; results may have been different with a "What I did last summer" type of topic. Furthermore, although a single topic type might be all that is needed in an operational program, that does not imply that a single topic is sufficient. Different topics, even within the same topic type, elicit slightly different performances, and the reliability of the total score increases as the number of topics sampled increases.

Separate scores for discourse-level and sentence-level skills do not appear to present any advantage over a single holistic score. Computer scoring of writing samples (Writers Workbench) provides data that appear to be potentially useful for assisting writing instruction and in the development of scoring systems, but it is not a substitute for holistic scoring based on human judgments.

Writing performance clearly differs across language groups, just as TOEFL performance differs across language groups. But there is no evidence that the writing samples unfairly discriminate against any group. Again, the careful topic selection procedures must be emphasized. Some of the topics rejected during the pilot testing did indeed appear to be discriminatory. Further research with criterion scores that were independent of the TOEFL would be needed to fully answer any questions of possible bias.

Research on the development of the writing process and the evaluation of the products of that process is in a period of discovery. Well-designed studies eventually will yield more substantive information about how the various parameters we have described influence writing and its assessment, and how these factors might influence students who represent different populations differentially. Until we can design writing assessment tools with more confidence, all individuals who apply writing measures to decisions regarding students should be cognizent of the numerous variables that condition the interpretation of their results. These investigations were an attempt to provide a model of the design features that should be incorporated in research on writing assessment methodology and programs.

Bibliography

American ?sychological Association. (1984, February). Join technical standards for educational and psychological testing, 4th Draft. Washington, DC: American Psychological Association, Office of Scientific Affairs.

Breland, H. M., & Jones, R. J. (1982). Perceptions of writing skill. (ETS Research Rep. No. 1982-47). Princeton, NJ: Educational Testing Service.

Bridgeman, B., & Carlson, S. (1983). Survey of academic writing tasks required of graduate and undergraduate foreign students. (ETS Research Rep, No. 1983-18). Princeton, NJ: Educational Testing Service.

Bridgeman, B., & Carlson, S. (1984). Survey of academic writing tasks. Written Communication, 1(2), 247-256.

Carlson, S., & Bridgeman, B. Testing ESL student writers. In K. Greenberg, H. Wiener, & R. Donovan (Eds.), Writing assessment: Issues and strategies. New York: Longman (in press).

Carlson, S., Bridgeman, B., Camp, R., & Waanders, J. Relationships of admissions test scores to writing performance of native and nonnative speakers of English. GRE General Report and ETS Research Report (in preparation).

Carrell, P. L. (1982). Cohesion is not coherence. TESOL Quarterly, 16(4), 479-488.

Cherry, L. L., Fox M. L., Frase L. T., Gingrich P. S., Keenan, S. A., & Macdonald, N. H. (1983). Computer aids for text analysis. Bell Laboratories Record, May/June.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement. Washington, DC: American Council on Education, pp. 443-507.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). Factors in judgments of writing ability. (ETS Research Bull. 1961-15). Princeton, NJ: Educational Testing Service.

Freedman, S. W. (1979). How characteristics of student essays influence teachers' evaluations. Journal of Educational Psychology, 71, 328-338.

French, J. W. (1962). Schools of thought in judging excellence of English themes. Princeton, NJ: Educational Testing Service.

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). The measurement of writing ability. New York: College Entrance Examination Board.

Greenberg, [illegible] (1983). Writing tasks and students' writing performance (1). In B. Kwalick, M. Silver, and V. Slaughter (Eds.), Selected Papers from the 1982 Conference 'New York Writes'." New York: Instructional Resource Center, Office of Academic Affairs, The City University of New York.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). Testing ESL composition: A practical approach. Rowley, MA: Newbury House.

Kaplan, R. B. (1966). Cultural thought patterns in inter-cultural education. Language Learning, 16, 1-20.

Kaplan, R. B. (1972). The anatomy of rhetoric: Prolegomena to a functional theory of rhetoric. In Language and the teacher: A series in applied linguistics, 8. Philadelphia, Pa.: The Center for Curriculum Development, Inc.

Kaplan, R. B. (1976). A further note on contrastive rhetoric. Communication Quarterly, 14(2), 12-19.

Kaplan, R. B. (1977). Contrastive rhetoric: Some hypotheses. ITL, 39-40, 61-72.

Kaplan, R. B. (1982). Contrastive rhetoric: Some implications for the writing process. In I. Pringle, A. Freedman, & J. Yalden (Eds.), Learning to write: First language, second language. London: Longman.

Keech, C. (1982, November). Designing prompts for holistic writing assessments: Knowledge from theory, research, and practice. Part II. Practices in designing writing test prompts: Analysis and recommendations. In L. Ruth (Project Director), Properties of writing tasks: A study of alternative procedures for holistic writing assessment. (Final Report NIE-G-80-0034). Berkeley, CA: Bay Area Writing Project, Graduate School of Education, University of California.

Kiefer, K. E., & Smith, C. R. (1984). Textual analysis with computers: Tests of Bell Laboratories' computer software. Forthcoming in Research in the Teaching of English.

Macdonald, N. H., Frase, L. T., Gingrich, P. S., & Keenan, S. A. (1982). The Writer's Workbench: Computer aids for text analysis. Educational Psychologist, 17(3), 172-179.

Pike, L. W. (1979). An evaluacion of alternative item formats for testing English as a foreign language. (TOEFL Research Rep.). Princeton, NJ: Educational Testing Service.

Pitcher, B., & Ra, J. B. (1967). The relationships between scores on the Test of English as a Foreign Language and the ratings of actual theme writing. (Statistical Rep. No. 67-9). Princeton, NJ: Educational Testing Service.

Purves, A. (1984, March 8). International perspectives on writing assessment. Papers presented at the National Testing Network in Writing: Second Annual Conference on Writing Assessment, Tallahassee, FL.

Ruth, L. (1982). Sources of knowledge for designing writing test prompts. Chapter I, Part I. In L. Ruth (Ed.), Properties of writing tasks: A study of alternative procedures for holistic writing assessment. (Final Report NIE-G-80-0034). Berkeley, CA: Bay Area Writing Project, Graduate School of Education, Univ rsity of California.

Ruth, L. (1982, November). Designing prompts for holistic writing assessments: Knowledge from theory, research, and practice. Part I: Sources of knowledge for designing writing test prompts. In L. Ruth (Project Director), Properties of writing tasks: A study of alternative procedures for holistic writing assessment. (Final Report NIE-G-80-0034). Berkeley, CA: Bay Area Writing Project, Graduate School of Education, University of California.

Shaughnessy, M. P. (1977). Errors and expectations. New York: Oxford University Press.

Smith, C. R., & Kiefer, K. (1982, April). Writer's Workbench: Computers and writing instruction. Paper presented at the Proceedings of the Future of Literacy Conference, University of Maryland, Baltimore, MD.

Thompson-Panos, K., & Thomas-Ruzic, M. (1983). The least you should know about Arabic: Implications for the ESL writing instructor. TESOL Quarterly, 17 (4), 609-623.

Test of English as a Foreign Language. (1983). TOEFL test and score manual. Princeton, NJ: Educational Testing Service.

Wiseman, S. (1949). The marking of English composition in grammar school selection. British Journal of Educational Psychology, 19, 200-209.

Abstract

Four writing samples were obtained from 638 applicants for admission to U.S. institutions as undergraduates or as graduate students in business, engineering, or social science. The applicants represented three major foreign language groups (Arabic, Chinese, and Spanish), plus a small sample of native English speakers. Two of the writing topics were of the compare and contrast type and the other two involved chart and graph interpretation. The writing samples were scored by 23 readers who are English as a second language specialists and 23 readers who are English writing experts. Each of the four writing samples was scored holistically, and during a separate rating session two of the papers written by each student were assigned separate scores for sentence-level and discourse-level skills. Representative subsamples of the papers also were scored analytically with the Writer's Workbench computer program and by graduate-level subject matter professors in engineering and the social sciences.

In addition to the writing sample scores, TOEFL scores were obtained for all students in the foreign sample. GRE General Test scores were obtained for students in the U.S. sample and for a subsample of students in the foreign sample. Students in the U.S. sample also took a multiple-choice measure of writing ability.

Among the key findings were the following: (1) holistic scores, discourse-level scores, and sentence-level scores were so closely related that the holistic score alone should be sufficient; (2) correlations among topics were as high across topic types as within topic types· (3) scores of ESL raters, English raters, and subject matter raters were all highly correlated, suggesting substantial agreement in the standards used; (4) correlations and factor analyses indicated that scores on the writing samples and TOEFL were highly related, but that each also was reliably measuring some aspect of English language proficiency that was not assessed by the other; and (5) correlations of holistic writing sample scores with scores on item types within the sections of the GRE General test yie!ded a pattern of relationships that was consistent with the relationships reported in other GRE stıdies.